

Deep Multi-Labeling based Unsupervised Person Re-Identification



By

Zarmeena

2018-NUST-MS-CS-08-274517

Supervisor

Dr. Muhammad Shahzad
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree
of Masters in Computer Science (MS CS)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(September 2021)

Approval

It is certified that the contents and form of the thesis entitled "Deep Multi-Labeling based Unsupervised Person Re-identification" submitted by ZARMEENA HASSAN have been found satisfactory for the requirement of the degree

Advisor : Dr. Muhammad Shahzad

Signature: M. SHAHZAD

Date: 26-Sep-2021

Committee Member 1: Dr. Qaiser Riaz

Signature: Qaiser Riaz

Date: 25-Sep-2021

Committee Member 2: Dr. Muhammad Moazam
Fraz

Signature: M. Moazam Fraz

Date: 28-Sep-2021

Committee Member 3: Dr. Muhammad Imran Malik

Signature: Imran Malik

Date: 27-Sep-2021

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Deep Multi-Labeling based Unsupervised Person Re-identification" written by ZARMEENA HASSAN, (Registration No 00000274517), of SEecs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____ *M. SHAHZAD* _____

Name of Advisor: Dr. Muhammad Shahzad _____

Date: _____ **26-Sep-2021** _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

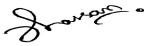
Dedication

I would like to dedicate this thesis to my parents for supporting my dreams in every possible way, my sister who has stood by my side since day one. I would like to thank some special people in my life, starting with my best friend Nimra; who is the literal definition of sincerity and epitome of pure talent, Mariam; who has been with me since the day I started my MS journey in NUST, Khadija; whose prayers for me are the reason I am here today and last but not the least Anum; who is the reason I have faith that good people exist in this world.

Certificate of Originality

I hereby declare that this submission titled "Deep Multi-Labeling based Unsupervised Person Re-identification" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEecs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEecs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: ZARMEENA HASSAN

Student Signature:  _____

Acknowledgment

I would like to thank my supervisor and my GEC members for guiding me and supporting me during my thesis process.

Table of Contents

1	Introduction and Motivation	1
1.1	Introduction	1
1.2	What is a Person Re-Identification?	1
1.2.1	Supervised Person re-Identification	2
1.2.2	Unsupervised Person re-Identification	2
1.3	Person Re-Identification Process	3
1.3.1	Detection	3
1.3.2	Extraction	3
1.3.3	Feature Representation	4
1.3.4	Matching	4
1.4	Person Re-Identification Applications	4
1.4.1	Visual Surveillance	4
1.4.2	Robotics	4
1.4.3	Content based video retrieval	4
1.5	Challenges	5
1.5.1	Pose Variation	5
1.5.2	Scale	5
1.5.3	Body Misalignment	6
1.5.4	Viewpoint Variation	6
1.5.5	Cross Domain	7
1.5.6	Low Resolution Image	8
1.5.7	High Visual Similarity Among Different People	9
1.5.8	Illumination Change	9
1.5.9	Occlusion	10
1.6	Person Re-ID scenarios	10
1.6.1	Single-shot vs Multi-shot	10
1.6.2	Open-world vs Closed-world	11
1.6.3	Long-term vs Short-term Re-ID	11
1.6.4	Small-scale vs Large-scale Re-ID	11
1.7	Motivation	11

1.8	Problem Statement	11
1.9	Solution Statement	12
1.10	Key Contributions	12
1.11	Upcoming Chapters	12
1.11.1	Literature Review	12
1.11.2	Design and Methodology	12
1.11.3	Experiments, Results and Analysis	12
1.11.4	Conclusion and Future Work	13
2	Literature Review	14
2.1	Unsupervised RE-ID	14
2.2	Unsupervised domain adaptation	16
3	Design and Methodology	18
3.1	Circle Loss based Model Learning	18
3.2	Multi-Label Learning	20
4	Experiments, Results & Analysis	23
4.1	Performance Evaluation	23
4.1.1	Rank-1 Accuracy	23
4.1.2	Rank-5 Accuracy	24
4.1.3	Rank-10 Accuracy	24
4.1.4	Mean Average Precision	24
4.2	Datasets	24
4.2.1	Market Dataset	24
4.2.2	MSMT17	24
4.2.3	Duke-MTMC-reID	25
4.3	Implementation Details	25
4.4	Quantitative Results	26
4.5	Comparison with State-of-the art Methods	27
4.6	Qualitative Results	29
4.6.1	Results on Market-1501	29
4.6.2	Results on Duke-MTMC-reID	30
5	Conclusion and Future Work	31
5.1	Conclusion	31
5.2	Future Work	31

List of Tables

4.1	Datasets used in the proposed Methodology	25
4.2	Results of proposed model on Market-1501 AND DukeMTMC- ReID	26
4.3	Results in comparison with State-of-the-Art Methods on Market- 1501	27
4.4	Results in comparison with State-of-the-Art Methods on Duke- MTMC-reID	27

List of Figures

1.1	Query(Test) Image and Gallery Image matching in Person Re-Identification	2
1.2	Re-Identification Pipeline	3
1.3	Pose Variation in Person Re-Identification.	5
1.4	Scale difference of same identity in Market Dataset.	6
1.5	Body Misalignment in Person Re-Identification.	6
1.6	An example of viewpoint variation scenario	7
1.7	Cross Domain in Person Re-Identification	7
1.8	Cross Domain Generalization via Deep Learning	8
1.9	Low Quality Data in Person Re-Identification.	8
1.10	Visual similarity in Person Re-Identification.	9
1.11	Illumination variance in Person Re-Identification.	9
1.12	Occlusion in Person Re-Identification.	10
3.1	Circle Loss based Model Learning	19
3.2	An overview of Multi-label Learning using Auxilliary Dataset .	21
4.1	Model Pipeline	26
4.2	mAP on Market-1501 w.r.t State of the Art Methods	28
4.3	mAP on Duke-MTMC-reID w.r.t State of the Art Methods . .	28
4.4	Accuracy on Market-1501	29
4.5	Mean Average Precision on Market-1501	29
4.6	Accuracy on Duke-MTMC-reID	30
4.7	Mean Average Precision on Duke-MTMC-reID	30

Abstract

Person re-identification is the problem of finding the given person in the non-overlapping cameras at a given time or finding the person that has appeared in the same camera but at different time stamps. The re-identification problem plays vital role in the field of visual surveillance. In real world scenario, the problem is not as easy as it seems because of the various obstacles including but not limited to occlusion, lighting, viewpoint variation and low-quality data. Mostly techniques are presented in supervised manner, but they seem to perform well only in the case of abundant labelled data which does not make sense in real life scenario because it is hard to label huge amount of data which increases in every passage of time and it requires a great amount of not only resources but time as well.

We present an unsupervised technique, that not only optimizes deep feature learning but also utilizes the significant information of auxiliary references other than visual feature similarity in large-scale re-Id.

Chapter 1

Introduction and Motivation

Person re-identification is the task of re-identifying a person that appeared in an non-overlapping camera at different time stamps. As simple as it sounds, the problem of re-identification has always been an challenging task due to various obstacles including but not limited to domain shift, occlusion, poor image quality and pose variation. Since, the person re-identification has become an important part in visual surveillance, now more than ever is a crucial need to develop something that caters the re-identification problem in real time scenarios.

1.1 Introduction

Person re-identification is a problem of pedestrian matching with relevant images. The problem typically required high number of data to train. The images for person re-identification consist of people in different lightning, poses, circumstances etc that makes the task more difficult to learn about the similar features.

1.2 What is a Person Re-Identification?

Person Re-identification is basically a task of matching a query(test) image to a gallery image. This task requires extensive training on large data to learn the features that are required for later matching. The problem that makes the task difficult is the images that are being matched are from non-overlapping cross camera that results the difference of pose, lightning, occlusion and as low quality of images.

We'll first discuss the two broader categories of person re-identification and then will be discussing the challenges one by one in the next sections

respectively.

In a broader category, the person re-identification can be done via two different methods of learning.

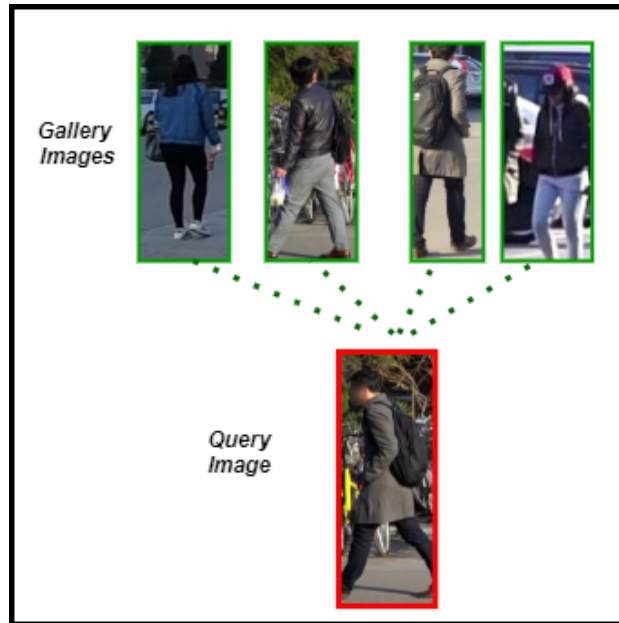


Figure 1.1: Query(Test) Image and Gallery Image matching in Person Re-Identification

1.2.1 Supervised Person re-Identification

Supervised Person re-identification is a method in which first the model is trained on a labelled training data and then after the model tries to match the test and gallery images accurately or in higher ranks. Even though supervised learning shows promising results, but in real time it doesn't make much sense to use supervised learning where we have abundant of unlabelled data. This type of learning requires higher number of data and extensive resources to do the training.

1.2.2 Unsupervised Person re-Identification

Unsupervised Person Re-identification does the task of pedestrian matching by training on the unlabelled source data. The learning can be done via different number of techniques including but not limited to cluster based

models, zero-shot learning etc. The unsupervised learning doesn't show phenomenal results because of the absence of labels, but it actually make sense with real life visual surveillance scenarios where there is extensive amount of data but unlabelled.

1.3 Person Re-Identification Process

The process of person re-identification starts from the localization of a person in a video data followed by the extraction of the target through bounding box followed by feature representation and then at last matching of query to gallery images. The following diagram represents the person re-identification thoroughly.

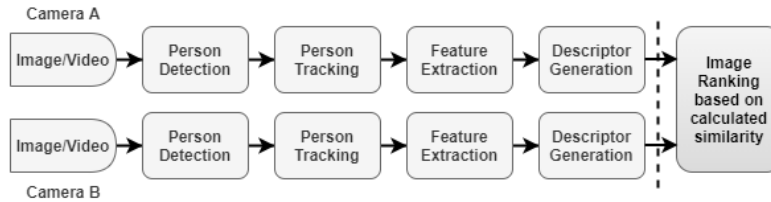


Figure 1.2: Re-Identification Pipeline

1.3.1 Detection

The first and foremost task in the process of person re-identification is the detection of the object. The surveillance camera footage contains hundreds of people and objects, and the main task is to detect the person image to perform further operations on the retrieved images.

1.3.2 Extraction

Once, the person is detected in the frames, the next step is to extract it from the raw footage or data. Because the matching is need to be done on the person image rather than the whole surrounding or redundant data. Hence the person images are extracted and hence the data is formed for further operations.

1.3.3 Feature Representation

This is the most crucial part of the process. The matching cannot be done on images because a system does not understand an image as an image but just an array of numbers representing colours. To perform the matching between the images, Features are extracted first and gallery to query matching is done via feature space.

1.3.4 Matching

This is the final step in the re-identification pipeline. the features obtained from query and gallery images are matched and based on that matching, the person is re-identified.

1.4 Person Re-Identification Applications

Person re-identification has been a hot topic in the computer vision domain because of its extensive use in important real life applications. The application include but not limited to video surveillance, robotics, and content based video creation.

1.4.1 Visual Surveillance

The first and foremost application of re-ID of the video surveillance.RE-ID can help in identifying the people who are involved in any kind of crime/misbehaviour. It was reported that between the month of January and march, 477 cars, 5982 cell phones and 1055 motorbikes were abducted in the Karachi, the largest city of Pakistan.

1.4.2 Robotics

As explained by Yuki et al. [1], while monitoring different visual fields, cameras are often mounted on robots to re-identify people using non-overlapping images and for that re-identification is widely utilized in the world of robotics.

1.4.3 Content based video retrieval

Person re-ID can be used in content based video retrieval because in the video retrieval process, interesting features and similarity measures among them play an important role and for that re-identification techniques can be utilised. [2]

1.5 Challenges

From Viewpoint variation to bad low resolution images, these obstacle make the task of re-identification quite difficult. We'll look into these issues in the following subsections.

1.5.1 Pose Variation

The pose variation can be the most challenging obstacle in the task re-identification. Because a person captured in a camera can appear with totally different pose in another non overlapping camera. This makes the job difficult because it'll be hard to find the significant similar feature representation in the both images. Figure 1.3 demonstrates the examples of pose variation in the data set.



Figure 1.3: Pose Variation in Person Re-Identification.

1.5.2 Scale

The scale issue in re-identification can be understood by the difference of scale of the person appeared across disjoint cameras. Hence the person appeared close to the camera will have clear features while the person captured from far always will have ambiguous and weak features. This is one of the common obstacles in the world of re-identification.



Figure 1.4: Scale difference of same identity in Market Dataset.

1.5.3 Body Misalignment

The body misalignment is the challenge where body parts of the same person are misaligned in terms of structure. This makes the task quite difficult as the misalignment causes the network to learn different feature embedding of the same person hence may cause mismatching.



Figure 1.5: Body Misalignment in Person Re-Identification.

1.5.4 Viewpoint Variation

Viewpoint variation is the challenge of the variation across non-overlapped cameras. For example in a camera, a person is watching towards the camera and in other camera, the person appears to be moving away from the camera. In this scenario, camera 1 will have more features and different features that

camera 2 which have weak features. Also, because of the view point variation, a pair of images can contain similar person with similar background, similar person with different background or different person with similar background. This problem has led to find those features which are invariant of the view-point variation and stand out across the disjoint cameras.



Figure 1.6: An example of viewpoint variation scenario

1.5.5 Cross Domain

This is the hottest problem where the model is trained in another data set and tested on different datasets. the datasets were obtained and formulated independent, so both the data set will have different environment, different type of clothing, may be different weather or different lighting etc. Hence, these models perform poor in these scenarios.



(a) Market-1501 Dataset

(b) MSMT17 Dataset

Figure 1.7: Cross Domain in Person Re-Identification

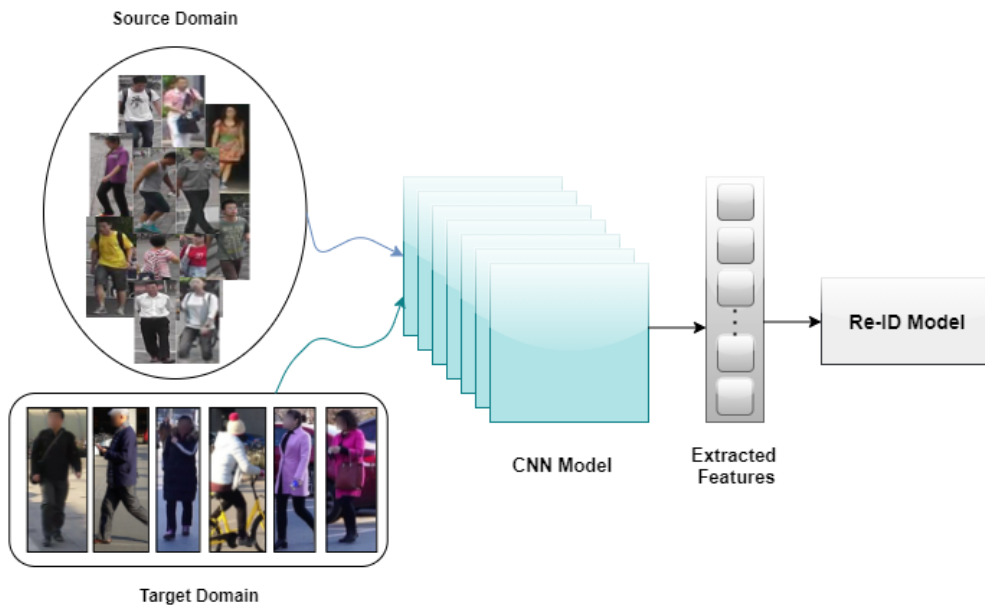


Figure 1.8: Cross Domain Generalization via Deep Learning

1.5.6 Low Resolution Image

Most of the data collected for the task of re-identification is based on surveillance footage. This results in low quality cctv frames and further operation on them (including bounding box generation) that makes the image of even more low quality that increases the loss of significant information.



Figure 1.9: Low Quality Data in Person Re-Identification.

1.5.7 High Visual Similarity Among Different People

One of the most prominent challenge in pedestrian challenges is the similar appearance of different people across non-overlapping cameras or within the same cameras. For example, in winters, mostly people wear the paddings or dark colour outfits that can be misunderstood as same person which results in wrong classification of the target image.



Figure 1.10: Visual similarity in Person Re-Identification.

1.5.8 Illumination Change

Illumination plays a critical role in the task. The images appeared in daylight or proper lighting contains more information than the images captured in poor lighting or night time. Not necessarily every place has good lighting and sometimes weather conditions can affect it and make it poor.



Figure 1.11: Illumination variance in Person Re-Identification.

1.5.9 Occlusion

Occlusion is the process of merging two or more objects hence making it difficult to locate or identify the target object. Occlusion often occurs in re-identification because the data is mainly obtained from the CCTV footage placed in densely populated areas for example university campus etc. Hence multiple objects appear in single frame and cause the target to be mingled with other objects making the process of re-identification more and more difficult.



Figure 1.12: Occlusion in Person Re-Identification.

1.6 Person Re-ID scenarios

Person re-identification is a broad term and can have different scenarios. The following subsection discuss four of the main scenarios in the domain of re-identification one by one.

1.6.1 Single-shot vs Multi-shot

Single shot, as the the name explains, has only single image against each query image to match. The data set is usually small whereas in Multi-shot, there are multiple images obtained from disjoint cameras to match against a test image. The data set is usually large in size and mostly used in video based- person re-identification

1.6.2 Open-world vs Closed-world

Open world refers to the real time scenario of re-identification where the images are coming live and you have to match them on the spot. The open world scenario is toughest as compared to Closed-world where the case is ideal and you have limited or pre-defined number of images to matche.

1.6.3 Long-term vs Short-term Re-ID

the long term re-identification is the matching of images that are taken with the gap of months or years.The scenario makes the process of re-identification difficult because conditions after months or years have significant impact on the appearance of a person.whereas the short term re-identification involves matching of images taken in the same day or few hours apart.

1.6.4 Small-scale vs Large-scale Re-ID

The scenario is said to be small-scale where the no of identities to be matches are less. In contrast to small-scale, the large scale re-identification involves matching over the span of large number of identities. The video based data sets come under the scenario of large scale r-identification

1.7 Motivation

To make video surveillance efficient is the main motivation behind this work. Even years of research on person re-identification has not been effective in real life scenarios mainly due to absence of extensive labelling the data or enormous amount of data captured on hourly basis. Even though few state of the arts works has been proven efficient but mainly they work on supervise manner, where as in real life, labels are not present. And if visual surveillance is done manually, it cannot be accurate and effective for maximum few hours due to the limitation of human body and human error. Thus a system is need to be made to tackle all the above mentioned issues.

1.8 Problem Statement

Supervised re-identification models faces scalability problem and because of that, unsupervised person re-identification has become center of attention domain due to its ability to tackle scalability problem of supervised re-id models. This work addresses the problem of learning discriminative features

of gallery-query pairs in the absence of labels across non-overlapping camera images.

1.9 Solution Statement

Developing a robust two step technique for solving the problem of extensive training on large data set that makes it easier to train and able to generalize well at the time of query to gallery matching.

1.10 Key Contributions

We propose a robust re-identification pipeline powered by implemented penalty loss followed by multi-label learning. The model is first trained on resnet-50 along side with data augmentation and penalty loss and then multi-label learning is done to acheive the discriminative features.

1.11 Upcoming Chapters

Later part of the thesis document is organized in the following chapters.

1.11.1 Literature Review

This chapter serves as a window into the notable work that has been done on person re-identification over the period of two decades. This section sets a research direction in this dissertation.

1.11.2 Design and Methodology

This chapter discusses our proposed person re-identification pipeline in detail. It breaks down our approach into different modules and provides an insight into their technical details.

1.11.3 Experiments, Results and Analysis

This chapter presents the experiments and their results. It also provides the analysis of the results in detail with handpicked examples.

1.11.4 Conclusion and Future Work

This chapter provides the final conclusive remarks and sheds light upon the future direction for the research community.

Chapter 2

Literature Review

The problem of Re-Identification in real time has caught the eyes of researchers in recent years. The Re-Identification problem can be divided into supervised, unsupervised and unsupervised domain adaptation. The upcoming literature discusses the above mentioned technique.

2.1 Unsupervised RE-ID

The work of Peixi et al. [3] presents an approach based on asymmetric dictionary learning based on multitasking. The method utilizes the labelled data set for the labeling of unlabeled data set. Hence making it unsupervised in nature.

Hehe et al. [4] proposed a technique based on initial clustering and then fine tuning of the convolutional neural network. The approach first make cluster on less amount of data and the gradually increase the circle of samples during fine tuning of the neural network.

Jingya et al [5] present a joint technique combining attribute-semantic learning and discriminative feature learning. The technique results in feature space that can be be assignable to any unseen data or unlabelled data.

The work of Weijian et al. [6] implements the idea of unsupervised learning between two different domains by initially transferring the labeled domain images and conserving their self similarity. It also assumes that transferred images and unlabeled or unseen data should not have any kind of similarity to them.

Zhun et al. [7] proposed the approach of homogeneous learning and heterogeneous learning in unsupervised domain adaptation. The approach assumes the camera invariance by mining positive pairs while it also implies the connection between the source and target domain by sampling training

pairs from both the datasets.

Hong-Xing et al. [8] demonstrates the efficiency of cluster based model in cross camera views. the model produces specific projection based on the clustering, (Here K-means model is used) of each of the image in cross camera.

Hong-Xing et al [9] presents an asymmetric metric learning for cross camera images. This deep learning based methodology eabvely relies on assymmetric metric learning which mitigate the feature distortion in each non overlapped disjoint camera views.

Elyor et al. [10] presented a unsupervised re-identification model to obtain the significant and discriminative features in cross-camera views by working on a dictionary based model.

Elyor et al. [11] proposed a joint sytem where model learns the feature representation and l1-norm based graph that is robust to outliers in re-identification

Rui et al. [12] tackles the obstacle of pose variation and large viewpoint in the domain of re-idnetification by incorporating patch matching in the image pairs by finding the significant and distinctive patches in the given data.

M. Farenzena et al. [13] proposes the algorithm to deal with the low resolution and viewpoint by combining the main aspects of human presentation that are consist of colours and patterns.

Hanxiao et al. [14] proposed an approach based on operations on foreground and background. The model not only extracts the significant patches from the foreground but it also tries to remove distortion from the background surrounding the target.

Yanbei et al. [15] presented a video based approach for re-identification. The methodology was build in an end-to-end manner and drive the best matching representation to each frame in inter and cross camera views by implementing margin based loss functions.

Minxian et al. [16] proposed a deep learning based framework to find the fundamental features within frames. The model first produces the human tracklets and then identify the association within the camera and correlation between the cross-camera of the tracklets in an end-to-end fashion.

The work of Sourya et al. [17] focuses on a subset- task of re-identification that is the extensive labelling of the data. The proposed methodology handles the given problem by choosing the optimal number of image-pairs to label and is based on k-partite graphs.

The work of Longhui et al. [18] is one of the most notable contributions to solve the challenge of re-identification. The work not only introduces the one of the largest re-ID dataset MSMT17 comprised of 4101 Identities captured through 15 non-overlapping cameras resulting in 126,441 bounding boxes but also provide a GAN based Network to cater domain gap.

Sinno et al [19] in his survey paper thoroughly explain the impact and importance of the relationship between transfer learning and various machine learning based approaches including but not limited to co-variate shift, Domain adaptation etc in the field of re-identification.

2.2 Unsupervised domain adaptation

The proposed approach of Mingsheng et al. [20] revolves around the problem of poor feature trasferability in the deep layers of neural network. The presented Deep Adaptation Network utilizes the reproducing kernel Hilbert space to embeds the hidden representation of task-specific or more precisely deep layers of the architecture.

Baochen et al. [21] proposed an statistical approach to solve the problem of domain adaptation between source and target data. In their work, they presented their model CORAL (CORrelation ALignmen) which reduces the domain shift by performing statistical operations on the training and testing distributions. The main advantage of this approach was its easiness to implement.

Baochen et al. [22] continued their work on unsupervised domain adaptation by presenting a Deep CORAL model which is an extension of their work on CORAL [21]. The presented model focus on the learning of nonlinear transformations in deep neural networks to coordinate the layer activations correlations.

The work of Rui et al. [23] deal with the main limitations of domain adversarial training. The proposed approach works on the Virtual Domain adaptation Model with the assumption that the domain containing high-density data should not be crossed by decision boundaries. Their work showed promising results on the digit and Wi-Fi recognition benchmarks.

The proposed technique Pietro et al. [24] deals the problem of unsupervised domain adaptation by minimizing the entropy between source and target domain. And to achieve this , the proposed model uses alignment along geodesics.

Eric et al. [25] proposed a framework to overcome the domain gap issue in re-identification The proposed model Adversarial Discriminative Domain Adaptation (ADDA) combines discriminative modeling along with the untied weight sharing and wraps the model with a GAN loss.

Yaroslav et al. [26] proposed a backpropogation based technique to minimize the domain shift gap between labelled source domain and unlabelled target domain. the proposed method can be implemented to any feed forward network by modifying the architecture slightly or adding a gradient reversal

layer along with few standard layers.

Mingsheng et al. [27] proposed a Joint Adaptation model which utilizes Adversarial training strategy to enhance the joint maximum mean discrepancy (JMMD) so that source distributions can be more distinct than target distributions.

Chapter 3

Design and Methodology

We introduced a technique to learn the significant feature embeddings in feature space using circle loss. For the proposed model, the network was trained on market dataset. for training, we used Resnet-50 as a backbone. The architecture produced 2048 dimension vector. Once the training is done, the Resnet-50 model is extracted and classification is layer is removed. Then we used soft multi-label learning, with assuming the following

- If two unlabelled images has less similarity, the pairs are similar.
- The images are compared to auxiliary dataset images, and if the images has less distance or more similarity with the reference images then the pair is said to be positive pair otherwise, it's negative.

Our proposed system has two main modules.

1. Circle Loss based Model Learning

The architecture in the presented model learn the feature embeddings using circle loss.

2. Multi-label Learning

The model does the multi-label learning using reference data on the unlabelled dataset.

This chapter discusses both these modules one by one in detail.

3.1 Circle Loss based Model Learning

The framework is consisted of an input data, a convolutional neural network, ending with a classification layer. We have used Market-1501 dataset as our

input dataset and as our backbone architecture, we have used Resnet-50. The circle loss [28] was added before the similarity score. To understand the working of circle loss, let's assume we have an anchor image and two image pairs consisting of positive and negative image each with respect to anchor image. Let's say the distance between positive and negative pair is same, unlike triplet loss which will find it ambiguous and treat them equally, circle loss will focus more on pushing the negative pair away as the positive image is already close, or pull the positive pair close rather than pushing the negative image away as it is already away from the anchor image, and for this unequal treatment towards the pairs, the model implies different level of importance on each image rather than treating them equally. We utilised the efficiency of the proposed loss function which will pull closer the images from similar class and pull away the images from different classes in reid.

Figure 3.1 demonstrates the concept in the form of a picture.

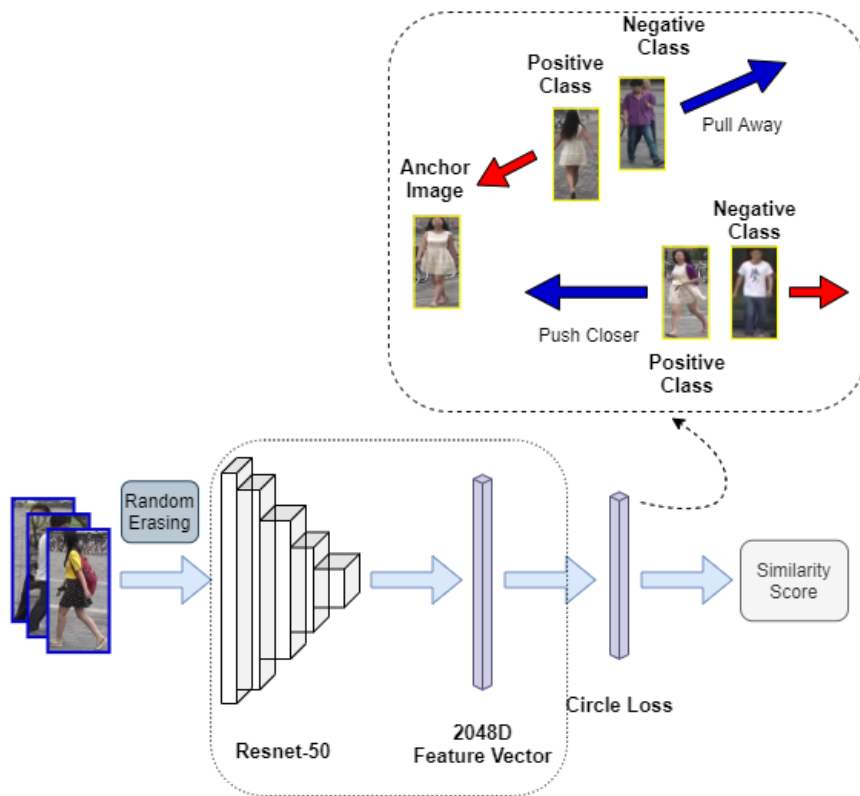


Figure 3.1: Overview of Circle-Loss based Model Learning: The backbone is based on ResNet-50 that generates the informative feature map. This feature map is forward to circle loss before predicting the similarity score

As explained earlier, the unified loss function can be described as follows

$$L_{uni} = \log\left[1 + \sum_{i=1}^K \sum_{j=1}^L \exp(\gamma(s_n^j - s_p^i + m))\right] \quad (3.1)$$

where γ is the scale factor and m is the margin.

And for class level labels, it will results to classification loss.

$$L_{am} = \log\left[1 + \sum_{j=1}^{N-1} \exp(\gamma(s_n^j + m)) \exp(-\gamma(s_p^i))\right] \quad (3.2)$$

And for a pair-wise label, it will degenerate to Triplet loss.

$$L_{tri} = \lim_{\gamma \rightarrow \infty} \frac{1}{\gamma} L_{uni} \quad (3.3)$$

$$L_{tri} = \max[s_n^j - s_p^i + m]_+ \quad (3.4)$$

So, the final circle loss can be described as follows:

$$L_{circle} = \log\left[1 + \sum_{j=1}^L \exp(\gamma \alpha_n^j (s_n^j - \Delta_n)) \sum_{j=1}^K \exp(-\gamma \alpha_p^i (s_p^i - \Delta_p))\right] \quad (3.5)$$

here Δ_p and Δ_n represents within class and between class margins respectively.

$$\begin{cases} \alpha_p^i = [O_p - s_p^i]_+ \\ \alpha_n^j = [s_n^j - O_n]_+ \end{cases} \quad (3.6)$$

3.2 Multi-Label Learning

As stated earlier, the multi-label learning works on two assumptions. To consider a pair positive, a pair should have high similarity and high similarity with reference agents as well. Firstly, The feature vectors of unlabelled data and reference data are obtained using the architecture explained in 3.1. Then the similarity between the unlabelled data images is calculated. For the similarity measure between the images, cosine similarity was used. To check the cross-camera pairs similarity, the model calculates the similarity between the unlabelled data and the reference labelled images. This process results in a multi-label vector where each entry represents a vote from reference agents. The multi-label learning utilizes the loss functions as presented in [29]

Figure 3.2 demonstrates the overview of the model.

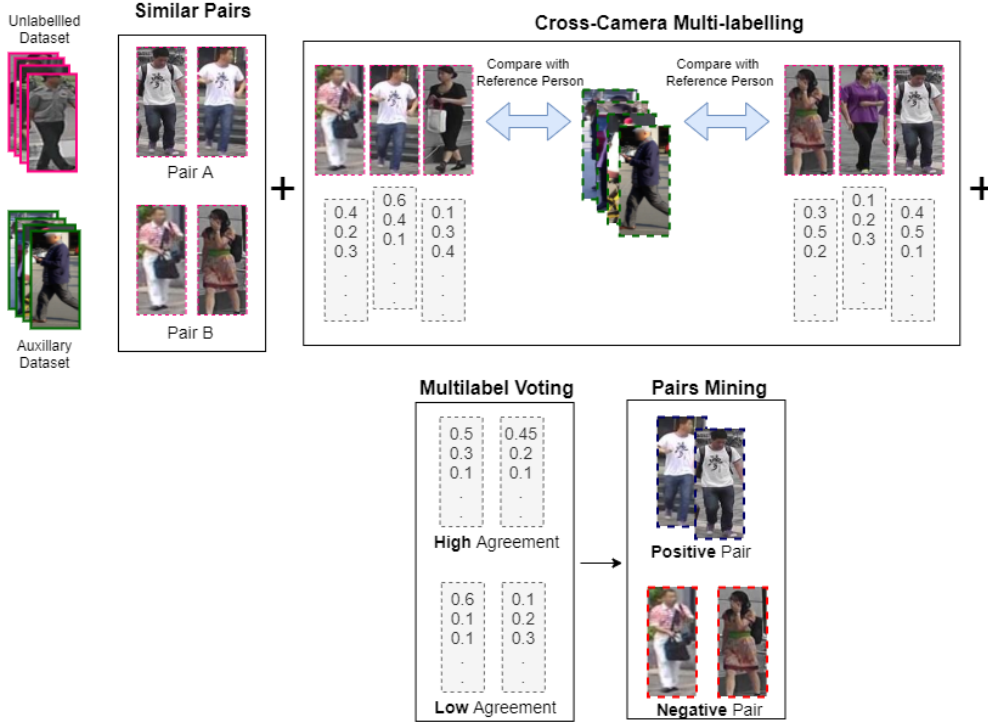


Figure 3.2: An overview of Multi-label Learning using Auxilliary Dataset

The multi-labelling comprises the following loss functions.

1. L_{RAL}

This loss basically ensures the that features are discriminative yet comparable with each other. The L_{AL} ensures that the reference images should be discriminative enough and should have representation power of associated images. Where L_{RJ} handles the reference comparability of reference images and unlabelled images. It implies that even if a reference image and an unlabelled image have visual similarity, they should have discriminative feature embeddings despite the dissimilar domain. So,

$$L_{AL} = \sum_k - \log l(f(z_k), (a_i))^{(w_k)} \quad (3.7)$$

and

$$L_{RJ} = \sum_i \sum_{j \in M_i} \sum_{\epsilon W_i} [m - \|a_i - f(x_j)\|_2^2]_+ + \|a_i - f(z_k)\|_2^2 \quad (3.8)$$

which concludes

$$L_{RAL} = L_{AL} + \beta L_{RJ} \quad (3.9)$$

2. L_{CML}

The L_{CML} is a Wasserstein distance based loss that implies that a person should have similar characteristics regardless that which camera was used to get the image. It should be independent of non-overlapping camera views. Hence,

$$L_{CML} = \sum_v \|\mu_v - \mu\|_2^2 + \|\sigma_v - \sigma\|_2^2 \quad (3.10)$$

3. L_{MDL}

Finally, to mine the positive/negative pairs based on the similarity threshold, the model defines L_{MDL} , where

$$L_{MDL} = -\log \frac{\bar{P}}{\bar{P} + \bar{N}} \quad (3.11)$$

The total multi-label learning loss will be as follow

$$L_{MAR} = L_{MDL} + \lambda_1 L_{CML} + \lambda_2 L_{RAL} \quad (3.12)$$

Chapter 4

Experiments, Results & Analysis

This chapter discusses the performance evaluation metrics, the experiments that have been carried out and their results. There is a detailed analysis based on the results as a guidance for the future work.

4.1 Performance Evaluation

Performance measures are put into place to evaluate the performance of the systems that are developed. For person re-identification, we have used r1, r5, r10 and mAP as our performance evaluation metrics. The r1 or precisely rank 1 accuracy describes the accurate re-identified pairs. Following sections describe these accuracy measures in details. The r5 accuracy tells that accurate pair is lied in top5 returned matches and same goes for r10 but in top 10 respectively. Mean Average Precision is calculated by taking the mean of Average Precision.

4.1.1 Rank-1 Accuracy

This is the standard of accuracy in the domain of re-identification. And it can be defined as the ratio of classified predictions to the actual true label. Rank-1 accuracy, informally also known as R1 accuracy can be obtained by dividing total number of correct predictions by the total number of data points in the dataset.

4.1.2 Rank-5 Accuracy

Rank-5 has a very similar calculation with rank-1 accuracy: but that slight difference is rather than only considering top 1 prediction from the classifier, It takes into consideration the top-5 predictions from the model.

4.1.3 Rank-10 Accuracy

Rank-10 has a very slight difference with rank-5 accuracy: but that slight difference is that rather than only considering top one or five prediction from the classifier, It considers the top ten predictions from the model.

4.1.4 Mean Average Precision

Mean Average Precision, also known as mAP is the most used accuracy measure in re-identification. It is the mean taken of average precision (AP) over all classes.

4.2 Datasets

The proposed methodology utilizes two of the most commonly available data sets: Market-1501, DukeMTMC-reID and MSMT17. The following subsections below will discuss the characteristics of each dataset in details

4.2.1 Market Dataset

The Market-1501 [30] dataset, as the name self explains, is comprised of 1501 identities and has total number of 32,668 images taken from non-overlapping cameras.

4.2.2 MSMT17

The MSMT17 dataset [18], also one of the largest re-identification datasets, contains 126,441 person images of 4101 unique identities. The dataset has been collected over the span of multiple days from non-overlapping 15 cameras rather than a single day hence introduces variation in clothing and weather conditions. That makes it a great choice as an auxiliary dataset because of the variation in the dataset.

4.2.3 Duke-MTMC-reID

The Duke-MTMC-reID [31] dataset contains 1404 unique identities, where total number of images are 36,411 captured from 9 non-overlapping cameras in Duke university campus. The dataset is no longer publicly available.

Table 4.1: Datasets used in the proposed Methodology

Dataset	No. of Identities	No. of Images	No. of Cameras
MSMT17	4101	126,441	15
Market-1501	1501	32,668	06
Duke-MTMC-reID	1404	36,411	9

4.3 Implementation Details

We have fine-tuned the process of feature learning on Resnet-50 using Market-1501. The first module was pre-trained obtained from [32]. The classification block was removed entirely and replaced with 2048 dimension feature vector layer and then retrained within module 2. The model in module 2 was trained for 20 epochs with learning rate 0.0002 and Batch-size of 200. The experiments were performed on NVIDIA-Tesla-P40 GPU that has 24GB Memory.

Figure 4.1 illustrates the complete pipeline of the detection algorithm

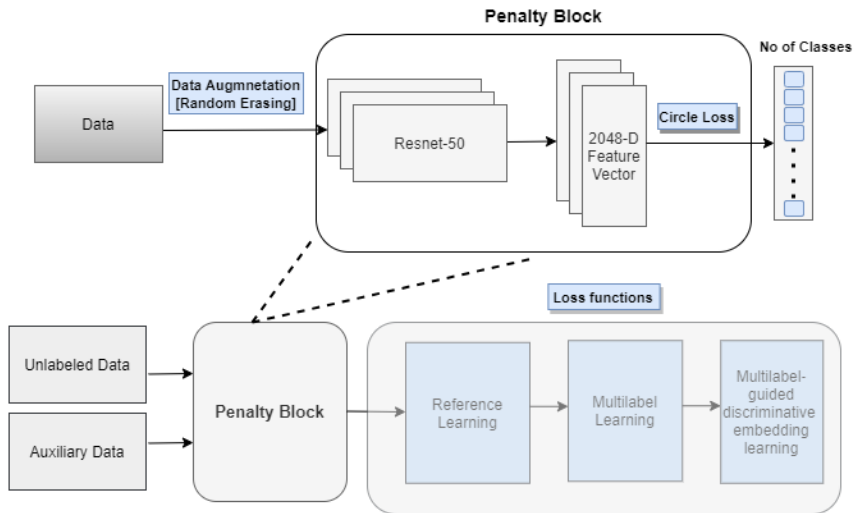


Figure 4.1: Model Pipeline

4.4 Quantitative Results

Table 4.2: Results of proposed model on Market-1501 AND DukeMTMC-ReID

Model	rank-1(%)	mAP	rank-1(%)	mAP
Ours	63.005	38.146	55.566	37.346

The proposed algorithm was performed on Market-1501 and DukeMTMC-reID. The table 4.2 shows the rank-1 accuracy and mean Average precision on mentioned datasets.

4.5 Comparison with State-of-the art Methods

Table 4.3: Results in comparison with State-of-the-Art Methods on Market-1501

Methods	Conference	rank-1(%)	rank-5(%)	mAP
TJ-AIDL	CVPR'18	58.2	74.8	26.5
PTGAN	CVPR'18	38.	57.3	15.7
SPGAN	CVPR'18	51.5	70.1	27.1
HHL	ECCV'18	62.2	78.8	31.4
DECAMEL	TPAMI'19	60.2	76.0	32.4
CamStyle	IEEE-TIP-19	58.8	78.2	27.4
CR+GAN+LMP	ICCV'19	64.5	79.8	33.2
UCDA-CCE	ICCV'19	60.4	-	30.9
MAR	CVPR'19	67.7	81.9	40.0
IDL+ACL+GAM	CoRR'20	63.9	78.8	35.7
STAR-DAC	Pattern Recognition-2022	67	80.6	33.9
Ours	This work	63.005	80.048	38.146

Table 4.4: Results in comparison with State-of-the-Art Methods on Duke-MTMC-reID

Methods	Conference	rank-1(%)	rank-5(%)	mAP
LOMO	CVPR'15	12.3	21.3	4.8
UDML	CVPR'16	18.5	31.4	7.3
CAMEL	ICCV'17	40.3	57.6	19.8
PUL	ToMM'18	30.0	43.4	16.4
PTGAN	CVPR'18	27.4	43.6	13.5
SPGAN	CVPR'18	41.1	56.6	22.3
HHL	ECCV'18	46.9	61.0	27.2
MAR	CVPR'19	67.1	79.8	48.0
IDL+ACL+GAM	CoRR'20	47.2	63.8	28.1
STAR-DAC	Pattern Recognition-2022	56.4	72.1	31.6
Ours	This work	56.957	71.768	38.515

The following charts compare the mean Average Precision of state of the art algorithms with the proposed model's. Fig 4.2 and Fig 4.3 shows the

comparison on Market-1501 dataset and Duke-MTMC dataset respectively.

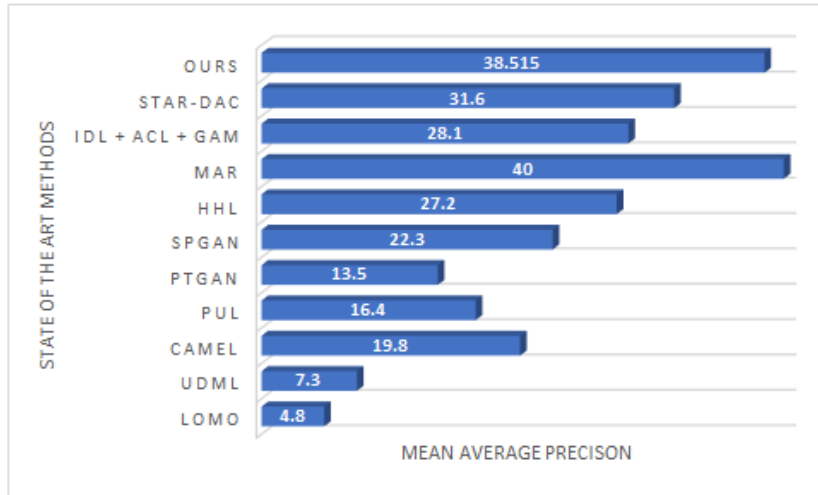


Figure 4.2: Comparison with other state-of-the-art methods over the years on Market-1501 dataset.

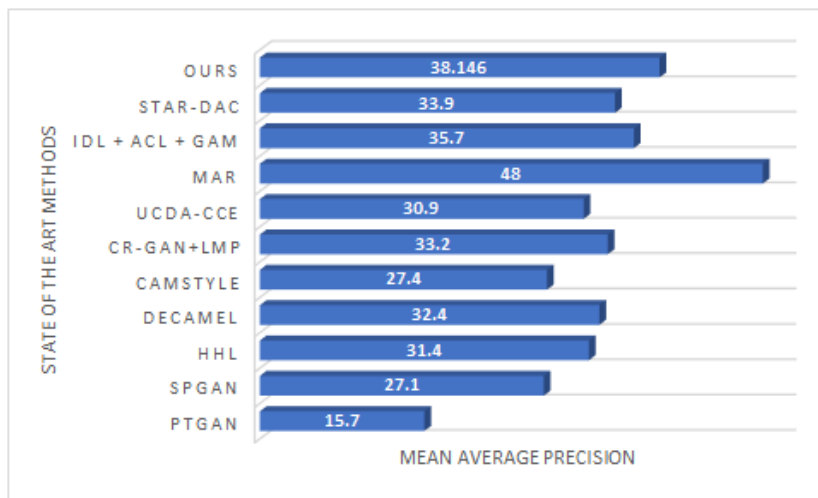


Figure 4.3: Comparison with other state-of-the-art methods over the years on Duke-MTMC dataset.

4.6 Qualitative Results

4.6.1 Results on Market-1501

Fig 4.4 and Fig 4.5 shows the Accuracy graph and mean Average Precision graph on Market-1501 respectively.

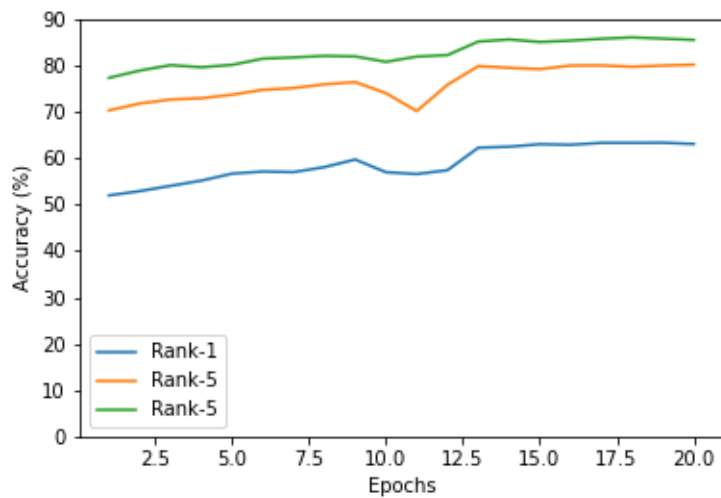


Figure 4.4: Accuracy on Market-1501

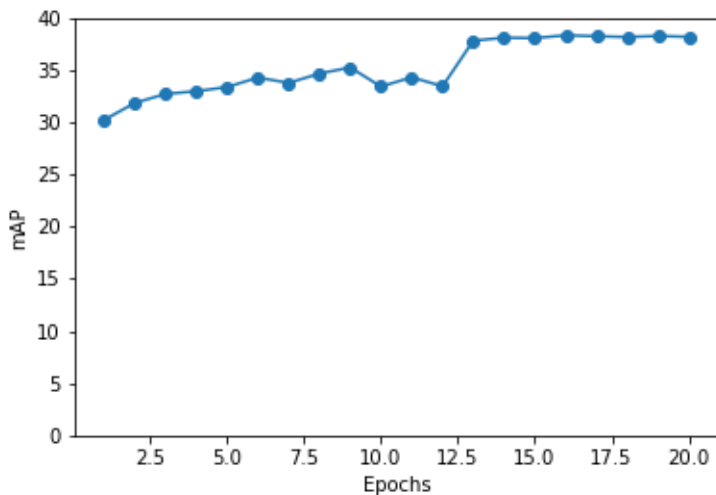


Figure 4.5: Mean Average Precision on Market-1501

4.6.2 Results on Duke-MTMC-reID

Fig 4.6 and Fig 4.7 shows the Accuracy graph and mean Average Precision graph on DukeMTMC-reID respectively.

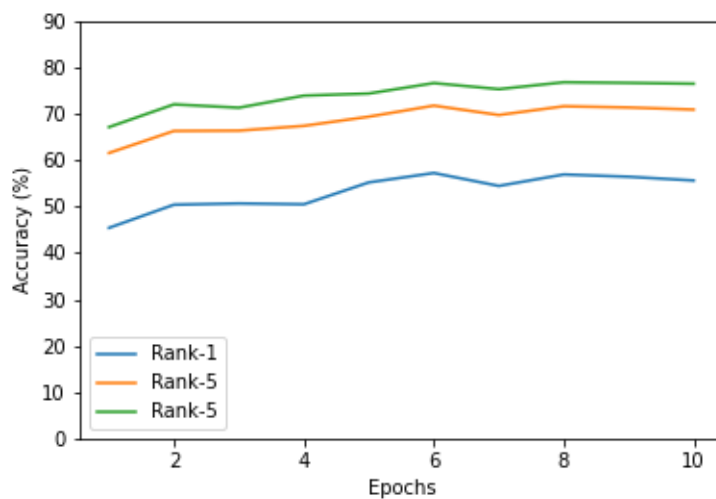


Figure 4.6: Accuracy on Duke-MTMC-reID

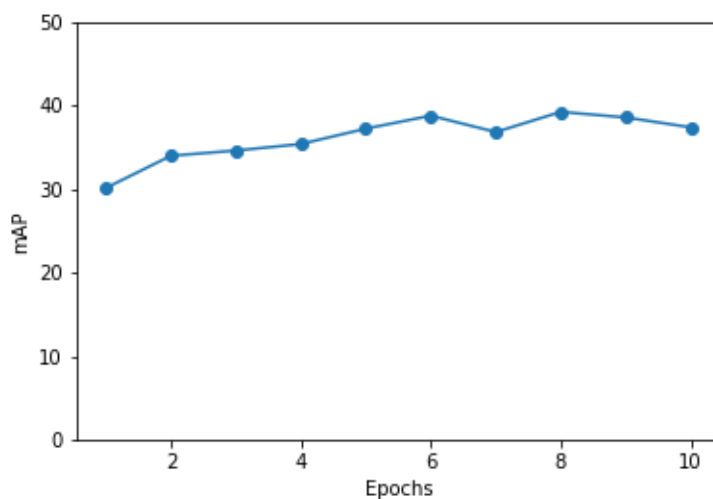


Figure 4.7: Mean Average Precision on Duke-MTMC-reID

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this research work, we present a two-step novel approach on feature extraction and then performing multi-label learning in unsupervised manner in person re-identification. First step is to train a network that is capable of finding distinct and unique features that can result in maximum query to gallery matching. To train the model for that, we've trained the Resnet-50 architecture on Market dataset along with circle loss. For Data Augmentation, the random erasing technique were applied in order to overcome the over fitting issue. As for the loss function, The circle loss was used to make the model more efficient in learning the Positive or Negative pairs which gives the final output.

In the next Module, the previously trained architecture was used in order to get the 2048D vector of the unlabelled and Auxiliary data set to perform the multi-label learning operation. We've utilised the effectiveness of soft multi-label learning [29] to demonstrate the effectiveness of our approach. In unsupervised Person re-identification, our approach shows the rank-1 accuracy of 63.008 percent and mAP of 38.146 on Market-1501 despite training on a relatively small data set. We will continue to focus on a single pipeline for unsupervised person re-identification.

5.2 Future Work

The current work evaluate its effectiveness on image based datasets. Its compatibility and performance on video based dataset can be checked in future. Also we have used Market dataset trained with circle loss. So, to further enhance the effectiveness of the proposed methodology, the dataset

can be varied for training.

Bibliography

- [1] Y. Murata and M. Atsumi, “Person re-identification for mobile robot using online transfer learning,” in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 977–981, 2018.
- [2] B. V. Patel and B. B. Meshram, “Content based video retrieval systems,” *CoRR*, vol. abs/1205.1641, 2012.
- [3] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, “Unsupervised cross-dataset transfer learning for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [4] H. Fan, L. Zheng, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, 05 2017.
- [5] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” *CoRR*, vol. abs/1803.09786, 2018.
- [6] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” *CoRR*, vol. abs/1711.07027, 2017.
- [7] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero- and homogeneously,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [8] H. Yu, A. Wu, and W. Zheng, “Cross-view asymmetric metric learning for unsupervised person re-identification,” *CoRR*, vol. abs/1708.08062, 2017.

- [9] H.-X. Yu, A. Wu, and W. Zheng, “Unsupervised person re-identification by deep asymmetric metric embedding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 956–973, 2020.
- [10] E. Kodirov, T. Xiang, and S. Gong, “Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification,” in *Proceedings of the British Machine Vision Conference (BMVC)* (X. Xie, M. W. Jones, , and G. K. L. Tam, eds.), pp. 44.1–44.12, BMVA Press, September 2015.
- [11] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, “Person re-identification by unsupervised ‘1 graph learning,” in *Computer Vision – ECCV*, 2016.
- [12] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586–3593, 2013.
- [13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2360–2367, 2010.
- [14] H. Wang, S. Gong, and T. Xiang, “Unsupervised learning of generative topic saliency for person re-identification,” in *Proceedings of the British Machine Vision Conference*, BMVA Press, 2014.
- [15] Y. Chen, X. Zhu, and S. Gong, “Deep association learning for unsupervised video person re-identification,” *CoRR*, vol. abs/1808.07301, 2018.
- [16] M. Li, X. Zhu, and S. Gong, “Unsupervised person re-identification by deep learning tracklet association,” *CoRR*, vol. abs/1809.02874, 2018.
- [17] S. Roy, S. Paul, N. E. Young, and A. K. Roy-Chowdhury, “Exploiting transitivity for learning person re-identification models on a budget,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7064–7072, 2018.
- [18] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer GAN to bridge domain gap for person re-identification,” *CoRR*, vol. abs/1711.08565, 2017.
- [19] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

- [20] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, p. 97–105, JMLR.org, 2015.
- [21] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” *CoRR*, vol. abs/1511.05547, 2015.
- [22] B. Sun and K. Saenko, “Deep CORAL: correlation alignment for deep domain adaptation,” *CoRR*, vol. abs/1607.01719, 2016.
- [23] R. Shu, H. H. Bui, H. Narui, and S. Ermon, “A dirt-t approach to unsupervised domain adaptation,” 2018.
- [24] P. Morerio, J. Cavazza, and V. Murino, “Minimal-entropy correlation alignment for unsupervised deep domain adaptation,” *CoRR*, vol. abs/1711.10288, 2017.
- [25] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” *CoRR*, vol. abs/1702.05464, 2017.
- [26] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by back-propagation,” in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1180–1189, PMLR, 07–09 Jul 2015.
- [27] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, p. 2208–2217, JMLR.org, 2017.
- [28] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, “Circle loss: A unified perspective of pair similarity optimization,” *CoRR*, vol. abs/2002.10857, 2020.
- [29] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, “Unsupervised person re-identification by soft multilabel learning,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Computer Vision, IEEE International Conference on*, 2015.

- [31] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *ECCV Workshops*, 2016.
- [32] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, “Joint discriminative and generative learning for person re-identification,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.